09:00-12:00

- Introduction to OrthoFinder and introduction to the study system and infrastructure
- Visualising OrthoFinder results

12:00-13:00 Lunch

13:00-14:00

- Spiderman and the adaptive story (dN/dS)

14:00-14:15 Break

14:15-16:00

- Spiderman and the adaptive story (dN/dS, cont.)

09:00-12:00

- Introduction to OrthoFinder and introduction to the study system and infrastructure
- Visualising OrthoFinder results

12:00-13:00 Lunch

13:00-14:00

- Spiderman and the adaptive story (dN/dS)

14:00-14:15 Break

14:15-16:00

- Spiderman and the adaptive story (dN/dS, cont.)

Genome annotation and comparative genomics

Part 2, afternoon

Teachers: Bram Danneels, Helle Tessand Baalsrud, José Cerca, Ole K. Tørresen

Oslo Bioinformatics Workshop Week 2023
11th December

WITH GREAT POWER COMES GREAT RESPONSIBILITY

- spiderman (or his uncle)

Spiderman and the Adaptive Story

- José Cerca

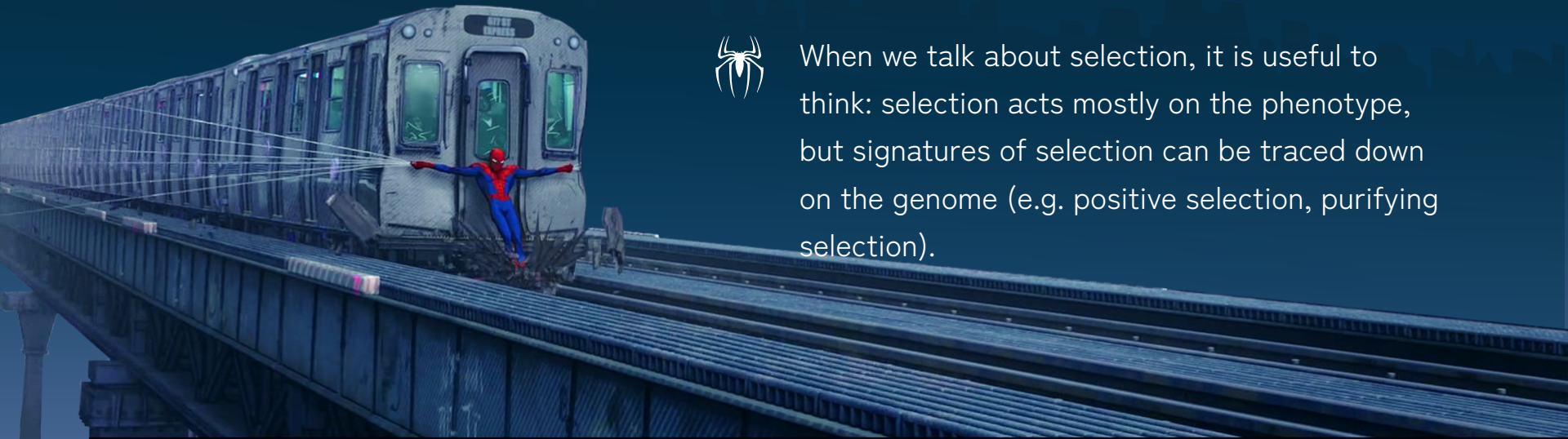
Molecular Evolution

One of the biggest goals of biology is to understand how adaptive pressures shape genomes (i.e. signatures of selection).

Molecular Evolution

One of the biggest goals of biology is to understand how adaptive pressures shape genomes (i.e. signatures of selection).

When we talk about selection, it is useful to think: selection acts mostly on the phenotype, but signatures of selection can be traced down on the genome (e.g. positive selection, purifying selection).

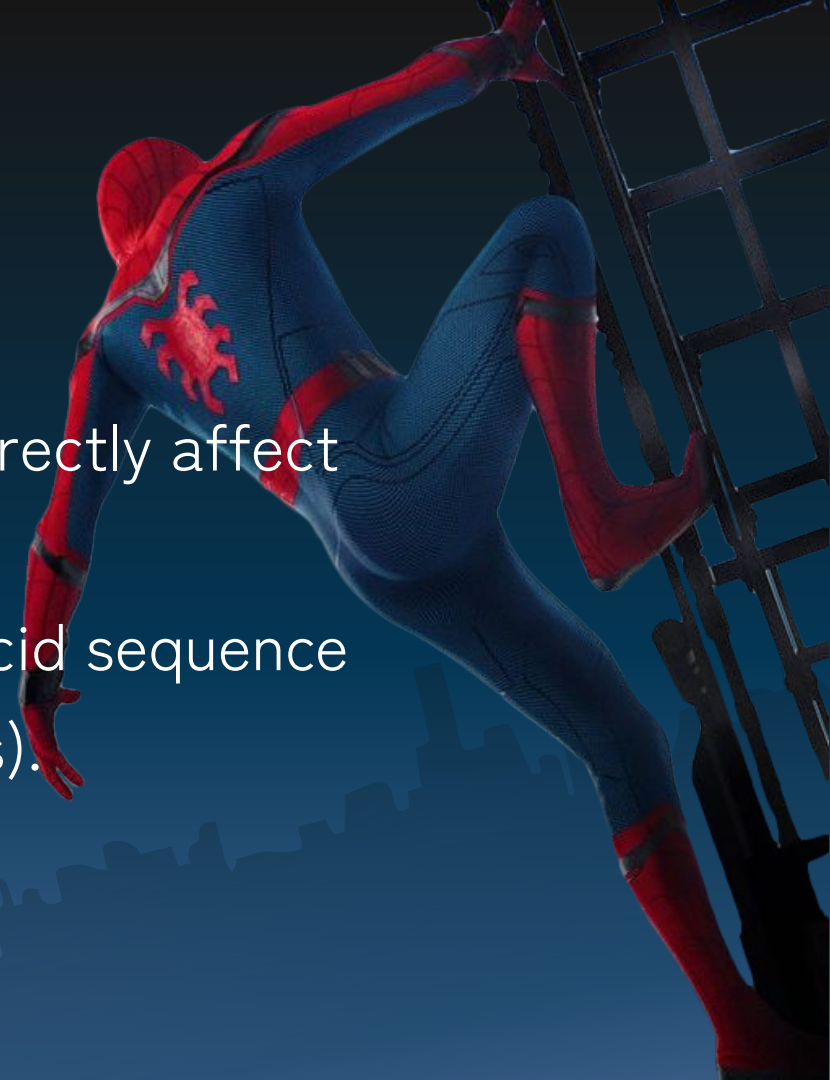


dN/dS

Also referred to Ka/Ks or ω

dN - nonsynonymous mutations can directly affect protein function

dS - mutations that leave the amino acid sequence unchanged (i.e. synonymous mutations).



dN/dS

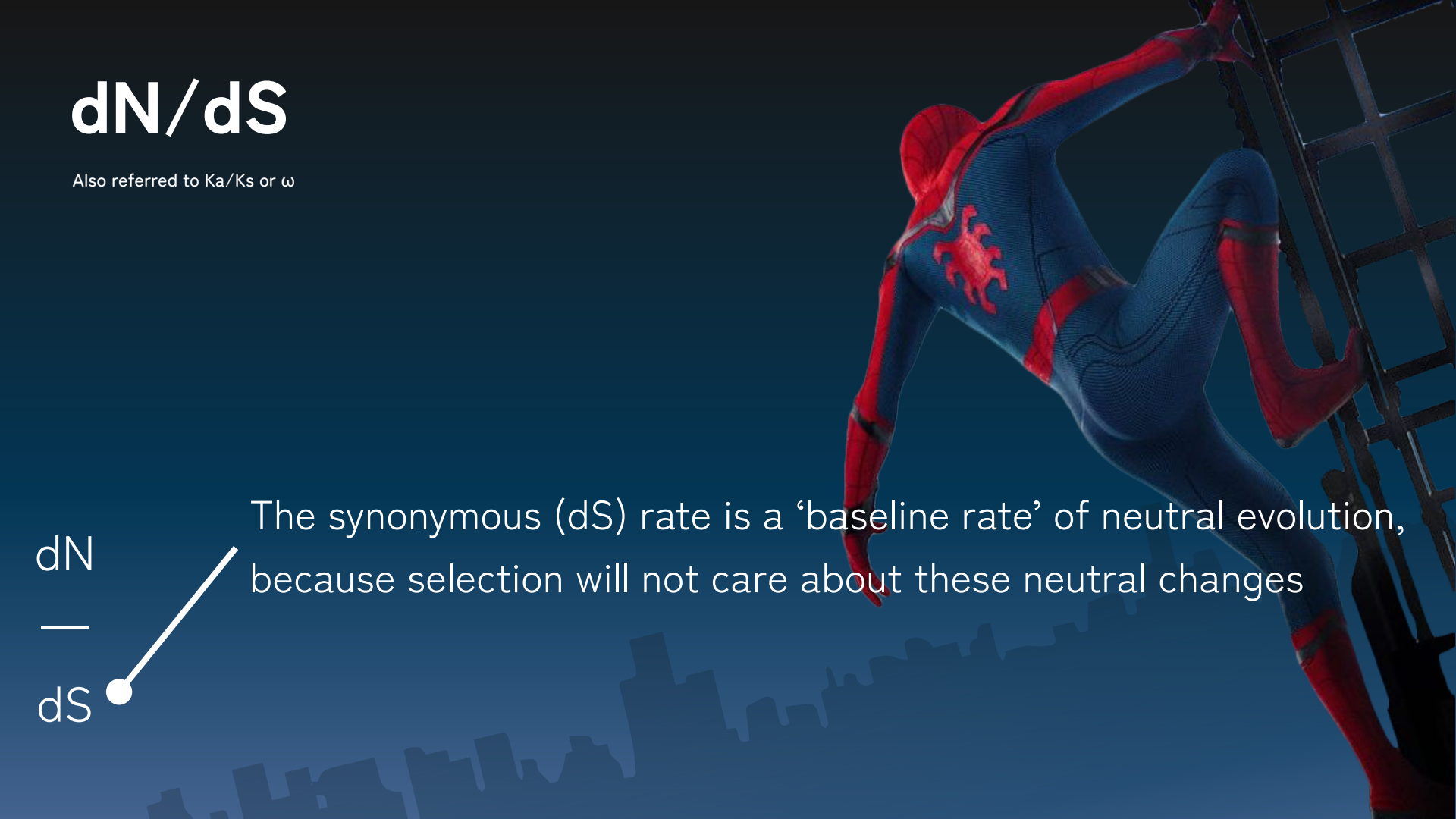
Also referred to Ka/Ks or ω

Ratio of non-synonymous (dN) to synonymous (dS) changes

$$\frac{dN}{dS}$$

dN/dS

Also referred to Ka/Ks or ω

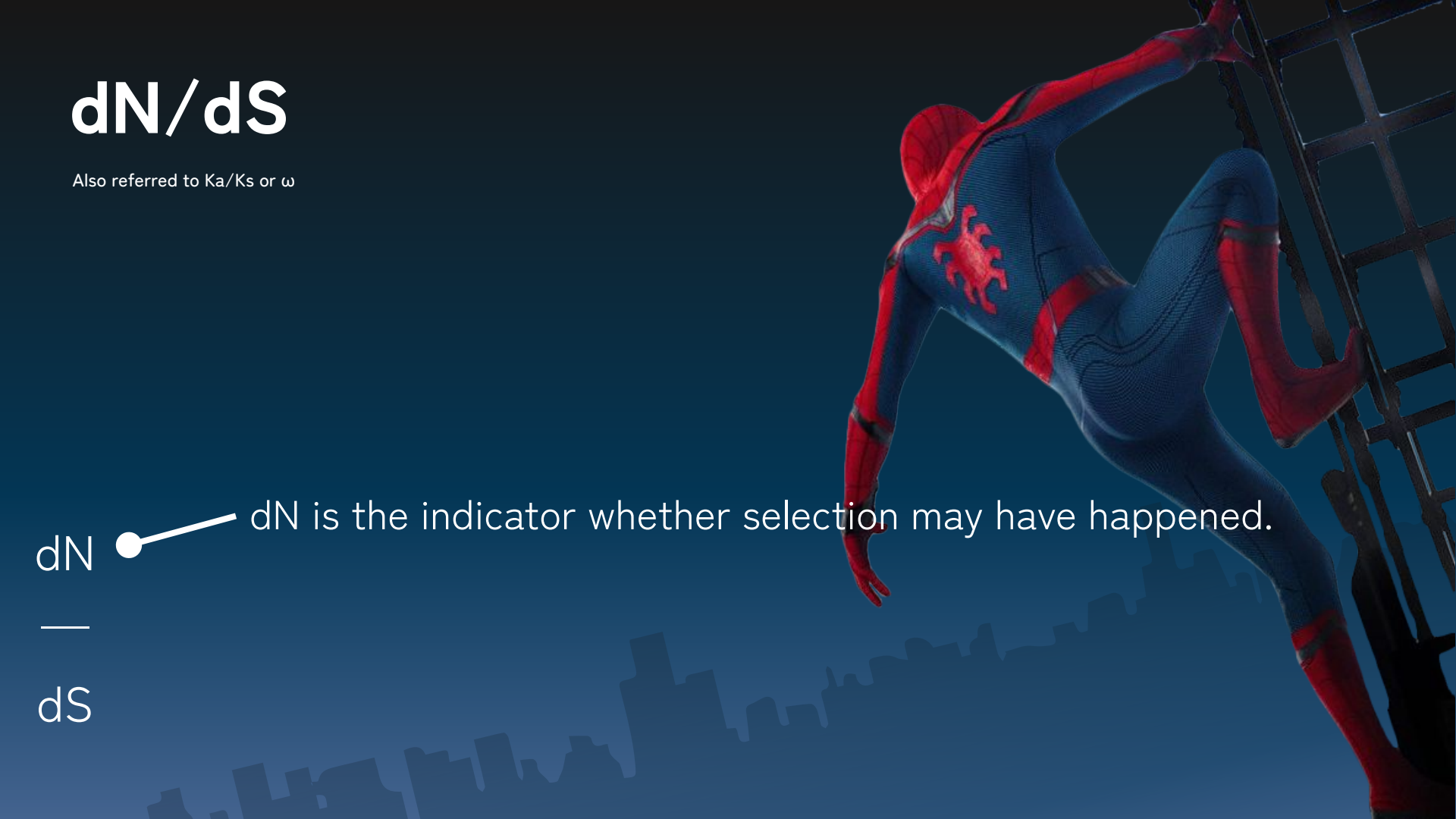


$$\frac{dN}{dS}$$

The synonymous (dS) rate is a ‘baseline rate’ of neutral evolution, because selection will not care about these neutral changes

dN/dS

Also referred to Ka/Ks or ω

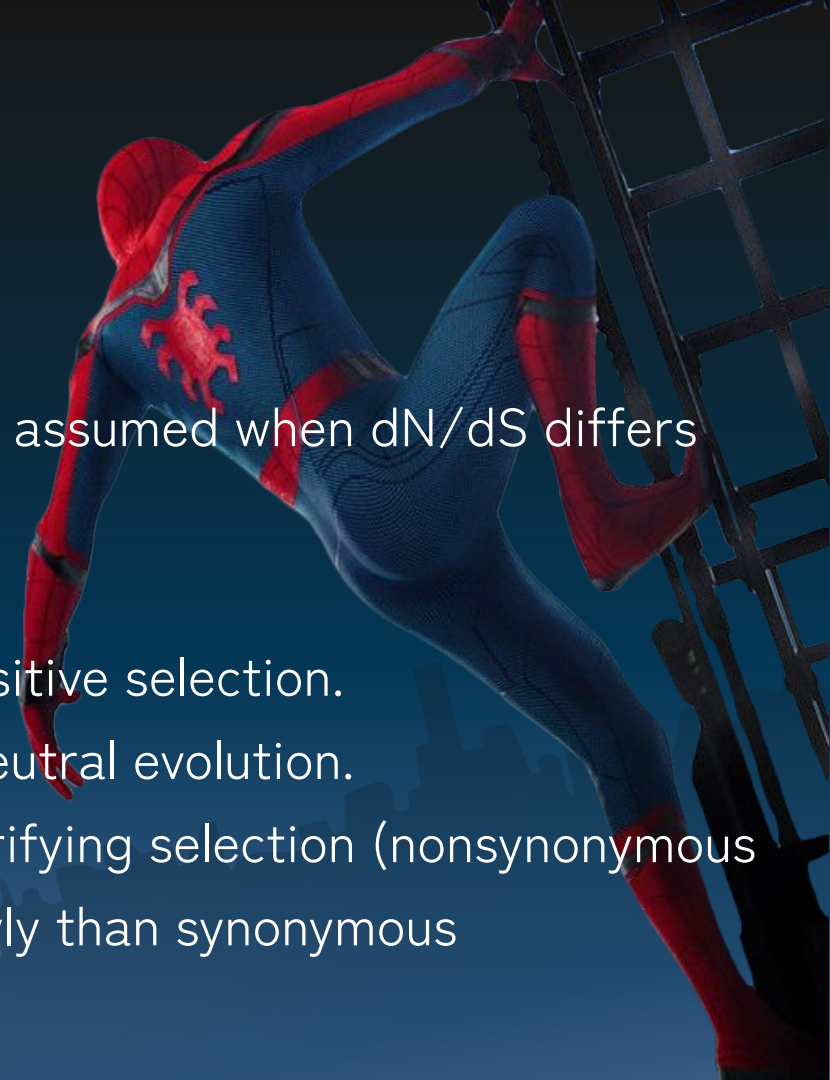


$$\frac{dN}{dS}$$

dN is the indicator whether selection may have happened.

dN/dS

Also referred to Ka/Ks or ω



The presence of selection is usually assumed when dN/dS differs from the neutral expectation.

$\frac{dN}{dS}$

When the **ratio is >1** we assume positive selection.
When the **ratio is ≈ 1**, we assume neutral evolution.
When the **ratio is <1** we assume purifying selection (nonsynonymous substitutions accumulate more slowly than synonymous substitutions)

HyPhy

Uses a phylogenetic tree and an alignment to understand molecular evolution.

HyPhy has tons of tests

BUSTED

Test for episodic **gene-wide selection using BUSTED** (Branch-site Unrestricted Statistical Test of Episodic Diversification).

RELAXED

Test for **relaxation of selection pressure** along a specified set of test branches using RELAX (a random effects test of selection relaxation)

ABSREL

Test for **lineage-specific evolution** using the branch-site method aBS-REL (Adaptive Branch-Site Random Effects Likelihood).

MEME

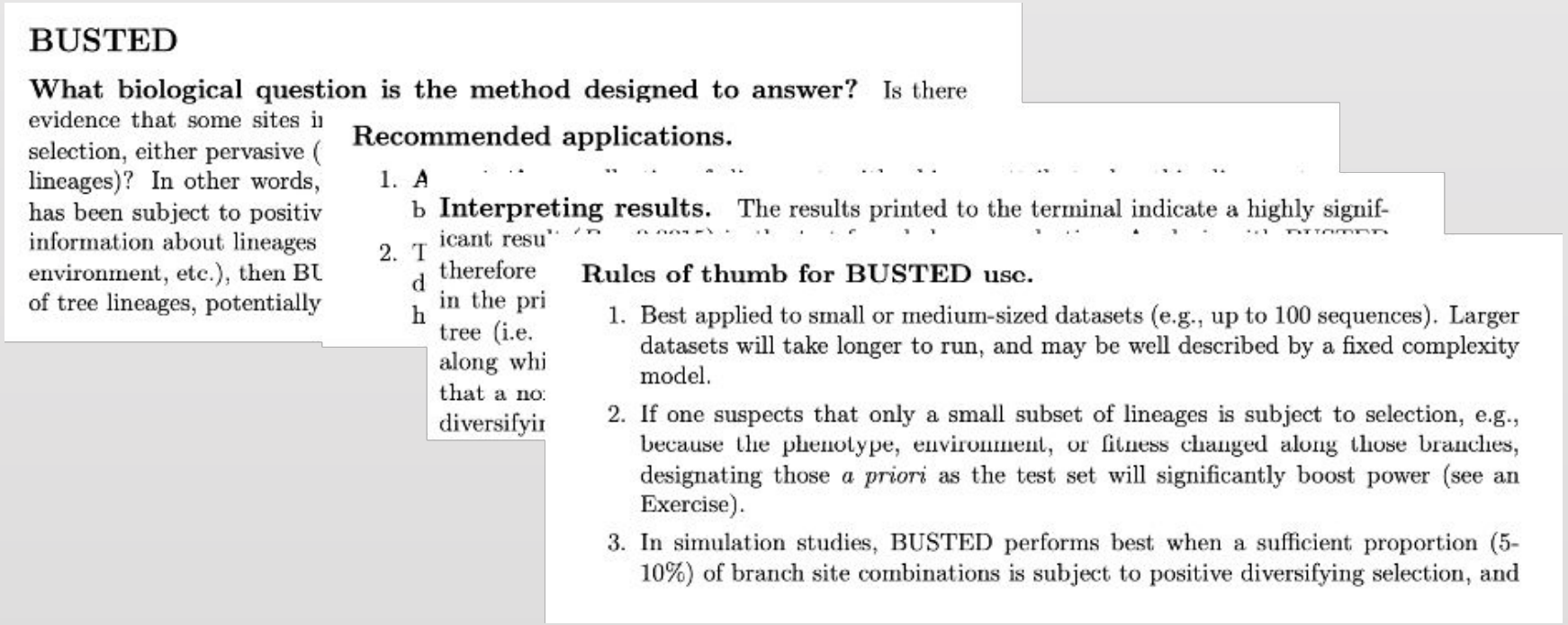
Test for **episodic site-level selection** using MEME (Mixed Effects Model of Evolution).

On the terminal, just try

hyphy -h

```
Available standard keyword analyses (located in /projects/ec146/miniconda3/envs/selection/share/hyphy/)
        meme    [MEME] Test for episodic site-level selection using MEME (Mixed Effects Model of Evolution).
        mh      Merge two datafiles by combining sites (horizontal merge).
        mv      Merge two datafiles by combining sequences (vertical merge).
        mcc     Compare mean within-clade branch length or pairwise divergence between two or more non-nested cladesd in a tree
        mclk    Test for the presence of a global molecular clock on the tree using its root (the resulting clock tree is unrooted, but one of the root branches can be divided in such a way as to enforce the clock).
        mgvsgy  Compare the fits of MG94 and GY94 models (crossed with an arbitrary nucleotide bias) on codon data.
        mt      Select an evolutionary model for nucleotide data, using the methods of 'ModelTest' - a program by David Posada and Keith Crandall.
        fel     [FEL] Test for pervasive site-level selection using FEL (Fixed Effects Likelihood).
        fubar   [FUBAR] Test for pervasive site-level selection using FUBAR (Fast Unconstrained Bayesian AppRoximation for inferring selection).
        fade    [FADE] Test a protein alignment for directional selection towards specific amino acids along a specified set of test branches using FADE (a FUBAR Approach to Directional Evolution).
        faa     Fit a multiple fitness class model to amino acid data.
        fmm      "Fit a model that permits double (and triple) instantaneous nucleotide substitutions"
        fst     Compute various measures of F_ST and (optionally) perform permutation tests.
        slac    [SLAC] Test for pervasive site-level selection using SLAC (Single Likelihood Ancestor Counting).
        sm      Perform a classic and structured Slatkin-Maddison test for the number migrations.
        sns     Parse a codon alignment for ambiguous codons and output a complete list/resolutions/syn and ns counts by sequence/position
        sw      Perform a sliding window analysis of sequence data.
        sa      Perform a phylogeny reconstuction for nucleotide, protein or codon data with user-selectable models using the method of sequential addition.
        sbl     Search an alignment for a single breakpoint.
```

The tutorial has a good overview of all the tests. We have no time to cover them all



BUSTED

What biological question is the method designed to answer? Is there evidence that some sites in selection, either pervasive (lineages)? In other words, has been subject to positiv information about lineages environment, etc.), then BU of tree lineages, potentially

Recommended applications.

1. A
   b
2. T
   d
   h

Interpreting results. The results printed to the terminal indicate a highly signif-
icant resu
therefore
in the pri
tree (i.e.
along whi
that a no
diversifyi

Rules of thumb for BUSTED use.

1. Best applied to small or medium-sized datasets (e.g., up to 100 sequences). Larger datasets will take longer to run, and may be well described by a fixed complexity model.
2. If one suspects that only a small subset of lineages is subject to selection, e.g., because the phenotype, environment, or fitness changed along those branches, designating those *a priori* as the test set will significantly boost power (see an Exercise).
3. In simulation studies, BUSTED performs best when a sufficient proportion (5-10%) of branch site combinations is subject to positive diversifying selection, and

Today we will be using

BUSTED

RELAXED

MEME

ABSREL

Test for **lineage-specific evolution** using the branch-site method aBS-REL (Adaptive Branch-Site Random Effects Likelihood).

Today we will be using

The input for the analysis is an aligned fasta file and a tree for that (we will be preparing this).

To get a fasta file with orthologs, you can use OrthoFinder to get orthogroups of cds regions (that's where we will start).

ABSREL

Test for lineage-specific evolution using the branch-site method aBS-REL (Adaptive Branch-Site Random Effects Likelihood).

Understanding what we will be doing

- ABSREL Determines whether a specific lineage or lineages have been subject to selection (“branch models”)

Understanding what we will be doing

- ABSREL Determines whether a specific lineage or lineages have been subject to selection (“branch models”)

- aBSREL asks whether some proportion of sites is subject to positive selection along specific branches or lineages of a phylogeny.

Understanding what we will be doing

- aBSREL Determines whether a specific lineage or lineages have been subject to selection (“branch models”)
- aBSREL asks whether some proportion of sites is subject to positive selection along specific branches or lineages of a phylogeny.
- aBSREL is recommended for
  - Exploratory testing for lineage-specific positive selection up to 100 sequences.
  - Testing of branches selected a priori.

Understanding what we will be doing

- aBSREL Determines whether a specific lineage or lineages have been subject to selection (“branch models”)
- aBSREL asks whether some proportion of sites is subject to positive selection along specific branches or lineages of a phylogeny.
- aBSREL is recommended for
  - Exploratory testing for lineage-specific positive selection up to 100 sequences.
  - Testing of branches selected a priori.
- If $\omega \leq 1$, the null model will have the same exact fit as the alternative model, and the resulting P-value is 1.
  Meaning, it doesn’t test for purifying selection.

Understanding what we will be doing

The test for lineage-specific diversifying selection is performed by comparing the full model versus the nested null model, and statistical significance is obtained by the likelihood ratio test.

aBSREL will correct all *P-values* obtained from individual tests for multiple comparisons using the Bonferroni-Holm procedures (targeting false positives).

Let’s get to it (45 min)

https://github.com/ebp-nor/genome_annotation_comparative_genomics_part2/blob/main/dNdS.md

Spiderman and the Adaptive Story

- PT 2

Zorro

```
(selection) [ec-josecer@login-1 ec-josecer]$ for i in 02_zorro/*zorro; do echo $i; awk '{ sum += $1 } END { if (NR > 0) print sum / NR }' $i; done
02_zorro/OG0000246.fai.zorro
7.6717
02_zorro/OG0000251.fai.zorro
7.48646
```

Both alignments had >7. My (arbitrary) cut-off is at 5.

Interpreting our results

Branch	Length	Rates	Max. dN/dS	Log(L)	AIC-c	Best AIC-c so far
0564_22	0.01	2	1.96 (52.27%)	-5402.41	11013.78	11009.72
0564_7	0.01	2	0.74 (5.19%)	-5402.40	11013.76	11009.72
Separator	0.01	2	197.32 (3.95%)	-5397.53	11004.02	11004.02
Separator	0.01	3	180.22 (4.08%)	-5397.53	11008.06	11004.02
0564_4	0.01	2	29.79 (2.15%)	-5394.37	11001.74	11001.74
0564_4	0.01	3	29.78 (2.15%)	-5394.37	11005.78	11001.74
0564_3	0.01	2	126.86 (3.14%)	-5388.59	10994.22	10994.22
0564_3	0.01	3	135.96 (3.05%)	-5388.59	10998.25	10994.22
0564_9	0.01	2	10.01 (8.61%)	-5388.37	10997.82	10994.22
...						
Node53	0.00	2	1.00 (100.00%)	-5371.63	10976.46	10971.76
0557_6	0.00	2	27.66 (100.00%)	-5371.32	10975.83	10971.76
0557_21	0.00	2	0.25 (1.96%)	-5371.30	10975.80	10971.76
0557_7	0.00	2	0.25 (1.96%)	-5371.30	10975.80	10971.76

The first table summarizes **model selection process**

- When two ω rates were assigned to branch Separator, this improved the AIC from 11009 to 11004. Three dN/dS rates got a worse AIC (aBSREL will then model two ω rates at the branch.

Interpreting our results

Branch	Rates	Max. dN/dS	Test LRT	Uncorrected p-value
0564_22	1	1.22 (100.00%)	0.11	0.43015
0564_7	1	0.61 (100.00%)	0.00	1.00000
Separator	2	197.72 (3.95%)	14.13	0.00029
0564_4	2	28.89 (2.15%)	4.81	0.03281
0564_3	2	127.66 (3.14%)	14.06	0.00030
0564_9	1	0.72 (100.00%)	0.00	1.00000
0564_1	1	1.07 (100.00%)	0.01	0.48208
...				
0557_21	1	1.00 (100.00%)	0.00	1.00000
0557_7	1	1.00 (100.00%)	0.00	1.00000

Adaptive branch site random effects likelihood test

Likelihood ratio test for episodic diversifying positive selection at Holm-Bonferroni corrected _p = 0.0500_
found **3** branches under selection among **44** tested.

* Node35, p-value = 0.00018
* Separator, p-value = 0.01251
* 0564_3, p-value = 0.01266

The second table shows the results of tests for episodic selection on individual branches.
At branch 0564_4 , the tested model included two ω rates, with the positive selection class taking on value 28.89 (2.15% proportion of the mixture).
Constraining this rate to range between 0 and 1 yields the likelihood ratio test statistic of 4.81, which maps to a P-value (before multiple test correction) of 0.03281.
Below the table we have corrected *p*-values.

Interpreting our results

```
* Node35, p-value =  0.00018
* Separator, p-value =  0.01251
* 0564_3, p-value =  0.01266
```

If a ‘Node’ shows under selection. You can see the json file and:

grep “Node3” *json
For example [show on itol]:
https://itol.embl.de/
(ANIA_04594,((GZUMIS1EN_004805_T1,GZUMRA1EN_001393_T1)Node3,GZUMVI1EN_006626_T1)Node2,(GGMOZO1EN_007347_T1,(GZPOHU1EN_000066_T1,GZMOAL1EN_008779_T1)Node9,GZLIHY1EN_002414_T1)Node7)


WITH GREAT POWER COMES GREAT
RESPONSIBILITY

WITH GREAT POWER COMES GREAT RESPONSIBILITY

Why did I start with this sentence?

WITH GREAT POWER COMES GREAT RESPONSIBILITY

I WANT TO MAKE A POINT

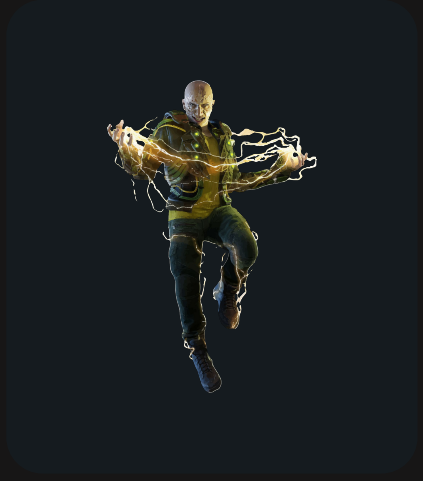
(WHICH I CONSIDER TO BE ONE OF THE MOST IMPORTANT)

dN/dS allows you answer some interesting questions

Is there evidence of selection operating on a gene?

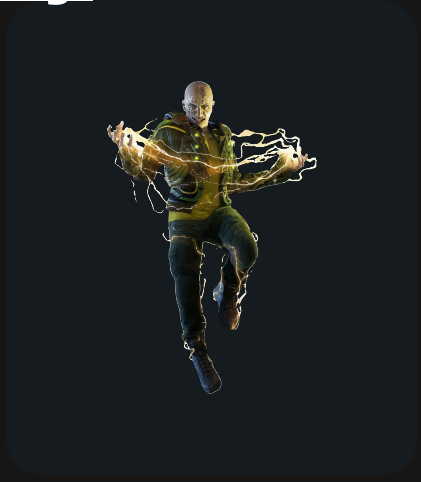
Where did selection happen?

When did selection happen? (internal phylogenetic branches)



What types of substitutions were selected for or against?

However, bear in mind that **with great power comes great responsibility**



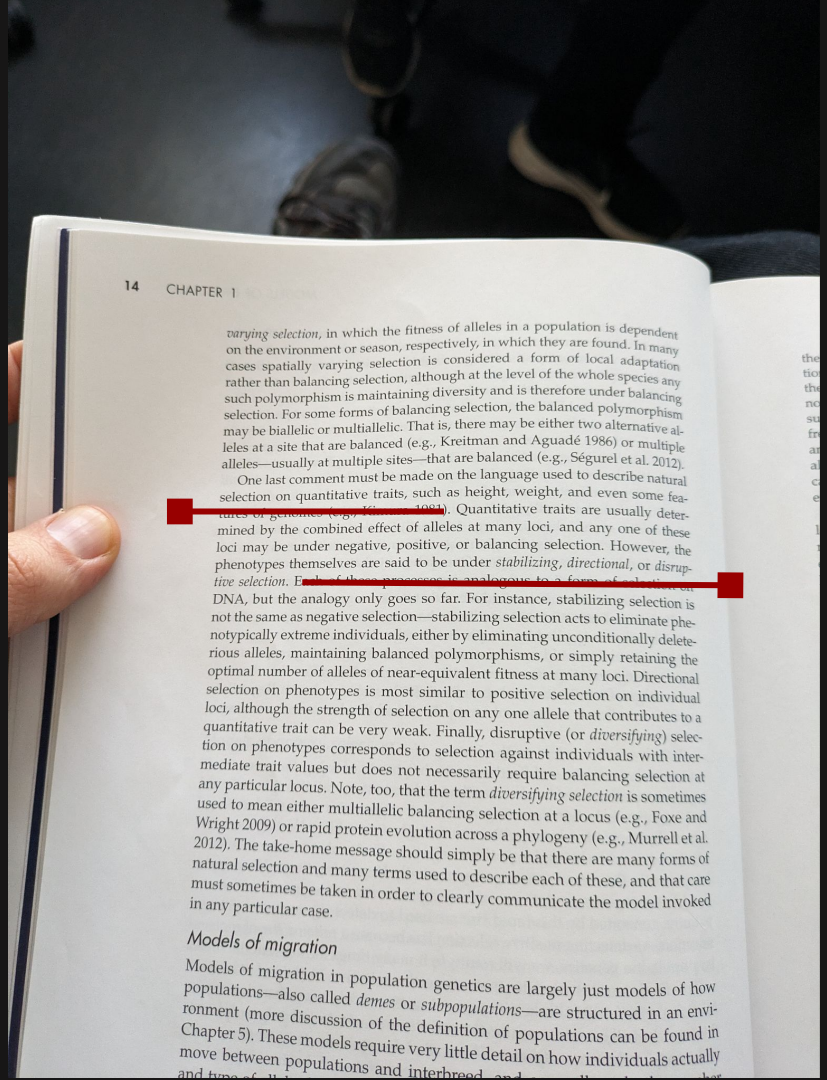
Consider

1) When we talk about selection it often gets “complicated”:

- Positive selection can mean diversifying selection / directional selection (See hyphy book 2007 for a discussion)

- Some forms of selection apply to phenotypes, but not genotypes (i.e. stabilizing selection, and I think directional selection too?)

Consider



14 CHAPTER 1

varying selection, in which the fitness of alleles in a population is dependent on the environment or season, respectively, in which they are found. In many cases spatially varying selection is considered a form of local adaptation rather than balancing selection, although at the level of the whole species any such polymorphism is maintaining diversity and is therefore under balancing selection. For some forms of balancing selection, the balanced polymorphism may be biallelic or multiallelic. That is, there may be either two alternative alleles at a site that are balanced (e.g., Kreitman and Aguadé 1986) or multiple alleles—usually at multiple sites—that are balanced (e.g., Ségurel et al. 2012).

One last comment must be made on the language used to describe natural selection on quantitative traits, such as height, weight, and even some features of genomes (e.g., [illegible]). Quantitative traits are usually determined by the combined effect of alleles at many loci, and any one of these loci may be under negative, positive, or balancing selection. However, the phenotypes themselves are said to be under *stabilizing*, *directional*, or *disruptive selection*. Each of these processes is analogous to a form of [illegible] DNA, but the analogy only goes so far. For instance, stabilizing selection is not the same as negative selection—stabilizing selection acts to eliminate phenotypically extreme individuals, either by eliminating unconditionally deleterious alleles, maintaining balanced polymorphisms, or simply retaining the optimal number of alleles of near-equivalent fitness at many loci. Directional selection on phenotypes is most similar to positive selection on individual loci, although the strength of selection on any one allele that contributes to a quantitative trait can be very weak. Finally, disruptive (or *diversifying*) selection on phenotypes corresponds to selection against individuals with intermediate trait values but does not necessarily require balancing selection at any particular locus. Note, too, that the term *diversifying selection* is sometimes used to mean either multiallelic balancing selection at a locus (e.g., Foxe and Wright 2009) or rapid protein evolution across a phylogeny (e.g., Murrell et al. 2012). The take-home message should simply be that there are many forms of natural selection and many terms used to describe each of these, and that care must sometimes be taken in order to clearly communicate the model invoked in any particular case.

Models of migration

Models of migration in population genetics are largely just models of how populations—also called *demes* or *subpopulations*—are structured in an environment (more discussion of the definition of populations can be found in Chapter 5). These models require very little detail on how individuals actually move between populations and interbreed, and [illegible]

Hahn, Molecular Population Genetics

Consider

2) Never assume a function of a gene **acritically.**

Hox genes have conserved functions. But tons of genes do not.

So, if you identify a gene using a model organism, and then you do positive selection scan, it can mean selection is acting but it can have a completely different function than you think.

Consider

2) Never assume a function of a gene **acritically.**

Moonlighting proteins: proteins that have one main, usually basic, cellular function, and another, usually more specialized, function in an unrelated process. Different parts of the proteins can be responsible for the different functions'

Neofunctionalization – when a gene gets a novel function

Genetic co-option the employment of conserved gene functions or pathways in a new process, for example the formation of new traits.

Review

Functional genomic tools for emerging model species

Erik Gudmundsdottir,[1,*] Christopher W. Wheat,[2] Abderrahman Khila,[1,3] and Arild Husby[1,*]

Most studies in the field of ecology and evolution aiming to connect genotype to phenotype rarely validate identified loci using functional tools. Recent developments in RNA interference (RNAi) and clustered regularly interspaced palindromic repeats (CRISPR)-Cas genome editing have dramatically increased the feasibility of functional validation. However, these methods come with specific challenges when applied to emerging model organisms, including limited spatial control of gene silencing, low knock-in efficiencies, and low throughput of functional validation. Moreover, many functional studies to date do not recapitulate ecologically relevant variation, and this limits their scope for deeper insights into evolutionary processes. We therefore argue that increased use of gene editing by allelic replacement through homology-directed repair (HDR) would greatly benefit the field of ecology and evolution.

Highlights

<<<< nice read on the subject.

Consider

3) Our field is heavily biased to some genes such as large-effect genes

THE QTN PROGRAM AND THE ALLELES THAT MATTER FOR EVOLUTION: ALL THAT'S GOLD DOES NOT GLITTER

Matthew V. Rockman[1,2]

[1]*Department of Biology and Center for Genomics and Systems Biology, New York University, 12 Waverly Place, New York, NY 10003*

[2]*E-mail: mrockman@nyu.edu*

Consider

4) It is hard to calculate recombination rates, and this can influence your results.
5) (poor)-alignments can influence your results
6) Comparing data with millions of years can influence your interpretations

THANK-YOU!

Hope you enjoyed "the story"

Evaluation form

Please fill out and give feedback!